Speech Sound Coherence within a Sentence Context

Elizabeth R. Troiano

The Ohio State University

Abstract

Best et al. (2008) found that selective attention was strengthened over time the longer participants listened to a string of numbers from the same spatial location. The authors claimed that spatial continuity leads to an increase in selective attention over time the longer a person listens to a continuous auditory stream. The reasoning for this is that the auditory features of the object at a specific location remain spatially constant in an otherwise complex auditory scene. As a result of this, the stream would become stronger and easier to organize. This claim of increased stream strength over time was tested using words in the context of sentences by creating an experiment in which /s/ phonemes were spatially separated from words they could potentially attach to in the sentence. I created pairs of sentences, one containing a target word like "start" toward the beginning and the other containing the same word toward the end. The sentences were presented in one of the participants' ears, and the /s/ of the key word was presented in the opposite ear. I instructed participants to repeat the sentences they heard and recorded whether they said the key words with or without the /s/. If participants reported the key words with the /s/, it indicated that the phonemes in the sentence were subject to intrusion from other sounds in the environment. I hypothesized that participants would be less likely to say the key words with the /s/ later in the sentences, because the build-up of a sentence would make it clearer which sounds belonged in the sentence and which did not. However, data analysis suggests that there is no overall difference between voiced responses to early and late target word positions. This means that people might have a baseline level of selective attention that remains unchanged throughout the progression of a spoken sentence.

Introduction

Verbal communication is something humans do every day with ease. However, the processes involved in grouping speech sounds together to form words is more complicated than most realize. Each word that individuals perceive is made up of several speech sounds, or phonemes, that are vastly different from each other. Despite this, people can seamlessly integrate them into words that make sense. For example, the phoneme /s/ is a fricative, meaning it has a continuous, static-like sound. However, the phoneme /p/ is a plosive, which is a combination of a complete block of airflow and a rapid release that creates a period of silence and then a burst of noise. The phoneme /a/ is a vowel, which means it has several continuous, resonant frequencies. Putting these three sounds together to form the word "spa" thus involves the perceptual combination of several distinct acoustic events. The theory of Auditory Scene Analysis (Bregman, 1990) was posed in order to explain how this perceptual grouping is accomplished in the brain.

Bregman's (1990) model of Auditory Scene Analysis provides a way to understand which sounds get grouped together in the auditory system and which sounds do not. Through this framework, he attempts to explain how meaningful units (like words, also known as auditory objects) are parsed from "streams," or persistent auditory stimuli (like an opera singer sustaining a note). Originally the theory was created for non-speech sounds, and so predicts that separating the sound of a violin out from the rest of the instruments in an orchestra works the same way as separating out one voice in a crowded room with lots of competing talkers. Over time, researchers applied the model to speech sounds as well, attempting to explain how it is possible to separate out two overlapping voices on the radio, for example. Bregman's model identifies two modes of perceptual organization of sounds. *Primitive segregation* organizes all sounds

based on their similar acoustic characteristics, whereas *schema-based segregation* organizes sounds based on the listener's prior knowledge.

Primitive segregation is the idea that sounds with vastly different acoustic characteristics (e.g. frequency, spatial location, pitch) are segregated, or not interpreted as belonging together, while sounds that have similar acoustic characteristics are integrated, or grouped together. Primitive segregation relies completely on the acoustic signal that the listener receives. Bregman (1990) makes the claim that it is innate, automatic, and unchangeable. In other words, people are born with this ability, and it is not something that happens with conscious effort or that can be overcome. In contrast, schema-based segregation groups together sounds that the listener has learned are related through conscious attention and association. It represents an advanced level of processing, involving previous exposure to and memory for patterns (including words in a language, musical sequences, and environmental sounds) which ultimately influences how the sounds in an acoustic signal are interpreted.

Schema-based segregation moves beyond the acoustic characteristics of sounds, instead focusing on each sound's relationship to other sounds and/or its relationship to the listener during the process of segregation. For example, listeners have mental representations of the way phonemes are coarticulated, meaning pronounced differently when preceded or followed by a different phoneme, such that they can predict after hearing one phoneme the phoneme that might come next. In this way, schema-based segregation makes the process of Auditory Scene Analysis more efficient, allowing the listener to base their interpretations of sounds on representations they have made from prior experience. Both primitive and schema-based segregation can operate concurrently, although schema-based segregation may be weighted more heavily depending on whether someone has extensive experience with a particular situation. Bregman states that

sounds that are considered related through the use of primitive grouping are organized regularly in the environment and have similar acoustic properties, like spatial location. This leads to the prediction that speakers will naturally group sounds together that come from the same spatial location.

Best, Ozmeral, Kopčo, and Shinn-Cunningham (2008) investigated how people are able to group speech sounds coming from different spatial locations; this is possible, although less intuitive, as Bregman's (1990) claim would suggest. In their study, participants listened to four loudspeakers arranged in a semi-circle which simultaneously played a sequence of numbers such that each number came from a different loudspeaker. The participants experienced one of several conditions: the target loudspeaker presenting the numbers either remained constant or switched with each number, and the target talker speaking the numbers either changed or remained constant with each number. Participants were asked to pay attention to either one loudspeaker for all four numbers or to a different loudspeaker for each of the four numbers, indicated by a flashing light. Participants were asked to report back the numbers they heard, and the authors found that performance was better on the task when participants were listening to one talker from one loudspeaker the whole time. The accuracy of participants increased from number to number when the presentation location was fixed, regardless of whether talker voice changed or remained constant. This means that selective attention to a specific location was strengthened the longer the participants listened to the string of numbers, whereas talker continuity was not as influential. The overall digit accuracy when location changed was lower than the fixed location, indicating that spatial continuity aided participants in performing the task. The authors thus claimed that spatial consistency leads to an increase in selective attention over time the longer a person listens to a stationary auditory location. This is because the auditory features of the object

within the location remain spatially constant in an otherwise complex and distracting auditory scene. As a result of this, the strength of the stream would increase, and the stream would become easier to organize. Best et al.'s findings suggest that cohesion of an auditory stream is stronger as it progresses, which was tested in the current experiments using casually produced sentences.

One caveat in interpreting the results is that Best et al. (2008) used digit strings instead of sentences. Strings of numbers do not contain semantic context, or the logical joining of several words to create intelligible meaning, as sentences do. This might not lead to generalizable interpretations of speech processing because speech in daily life includes semantic context which may aid in interpretation when lots of other sounds (speech or otherwise) are present at the same time. To be sure that their result of selective attention being refined over time applies to daily speech processing, the current experiments extend Best et al.'s study to a more naturalistic environment by using sentence frames. The use of semantic (and therefore linguistic) context allows me to extend the results of Best et al.'s (2008) experiment to an explanation of the cocktail party effect (Cherry, 1953). The cocktail party effect refers to people's ability to focus on one talker in a loud room with lots of other competing talkers. Best et al. (2008) claimed that selective attention to a particular location strengthens with time, using participants' attention to a randomly presented series of numbers to demonstrate it. Their results could be applied to explain the cocktail party effect, because it might indicate that people are able to better focus on one specific person as a result of a refinement in selective attention to one location over time. In order to expand Best et al.'s claim in a more natural environment, sentence stimuli were used to explore the effects of context on the strengthening of selective attention over time.

Experiment 1

To test whether selective attention is strengthened as a sentence progresses, the amount of integration of extraneous sounds into sentences was measured at different time points. More specifically, a sentence was presented in one ear and an /s/ isolated to the opposite ear, asking participants to repeat the sentence and ignore any other sounds, attending only to one side. A base target word with which an /s/ sound could logically bind was positioned either at the beginning or towards the end of the sentence (Figure 1). For example, the sentence "The bare shelf was in the red cabinet" was presented in one ear, and an /s/ was presented in the opposite ear right before the word "bare" such that participants could perceive "spare shelf," if they integrated the extraneous sound. Both of the possibilities, "spare" and "bare," were judged to be equally likely within the sentence contexts by both the experimenter and the results of a pilot study. Segregation was measured by the number of times participants reported a voiced target word (the base: "bare") when the /s/ was lateralized to the opposite ear from the ear of presentation of the sentence. I hypothesized that late position sentences would have more segregation of the /s/ from the target word than early position sentences, indicating that strengthening selective attention over time leads to an increase in primitive segregation. Additionally, it would provide naturalistic support for Best et al.'s (2008) conclusion of a refinement in selective attention over time, showing that spatial continuity leads to an increase in primitive segregation.

a) b)

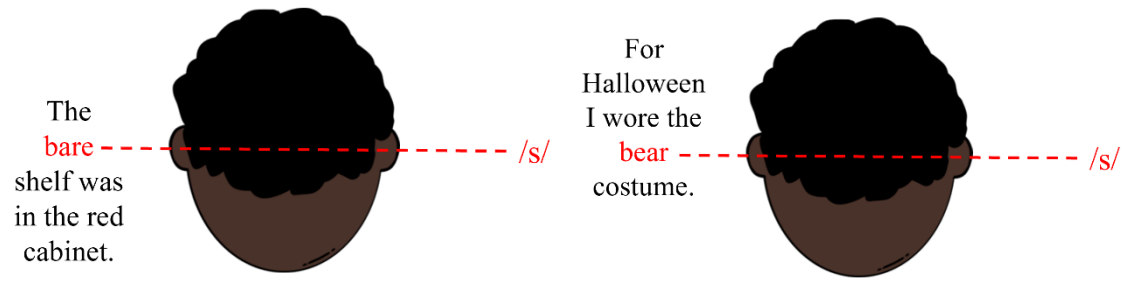


Figure 1: An example of the early and late position sentences are included in a) and b). The sentences are heard in the left ear with the target words highlighted in red. The /s/ of each target word is lateralized to the right ear and is presented at the same time it would be if it were attached to the rest of the base. The clipart used is by Lany (2015).

Method

Participants. Thirty-two Ohio State University undergraduate students (9 males, *M* = 18.56 years, range = 18–26 years) participated in the experiment and received course credit. Of those, four were excluded for being nonnative English speakers. This left a total of 28 participants who were self-reported native English speakers with normal hearing.

Stimuli. Sixteen monosyllabic target words starting with /s/ were selected such that when the /s/ was removed, the resultant base was a real word. For example, if the /s/ is removed from “spare,” the word “bare” is perceived. This is because, when producing the word “spare,” the aspiration (puff of air) produced when a /p/ is not preceded by a fricative is absent, making the word acoustically sound more like “sbare.” Despite this, English speakers perceive it as “spare”

because an /sb/ combination does not exist in the phonology of English; that is, the rules of English prohibit the combination of /s/ and /b/ phonemes. If the /s/ were separated to a different spatial location from the base, it could be grouped either with the rest of the base (resulting in a "spare" percept) or kept separate (resulting in a "bare" percept). If the word "bare" is reported, it was considered a *voiced* response, because /b/ is a voiced phoneme, whereas "spare" would be considered a *voiceless* response because /p/ is a voiceless phoneme. Voiced responses indicate that the /s/ is not considered part of the target word, and that the /s/ has been segregated from the rest of the sentence. Voiceless responses indicate that the /s/ and the base are instead being integrated together to create the perception of one word.

Sentences with and without the /s/ were created for the target words and these target words occurred either after one syllable (early position) or after seven to ten syllables (late position, M = 8.13 syllables) within the sentence. Each target word, for example "spit," would have both an early position sentence and a late position sentence. The paired sentences for each target word (early and late position) contained the same number of syllables in order to balance sentence length. These sentences were also designed to be semantically coherent when the target words occurred both with and without the /s/. An early sentence example would be, "He *spit/bit* away the rotten part of the peach," and a late position sentence would be, "The fight was so fierce that he *spit/bit* back his words." The list of target words and sentence frames is shown in Appendices A and B.

A female native English speaker recorded all sentence frames (32 sentences total) with the /s/+base word embedded. The target /s/ onset and offset in each of the sentences was marked using the sound editing software Praat (Boersma & Weenink, 2019), and the target word's /s/ was then removed, creating a second version of each of the 32 sentences without /s/. The length

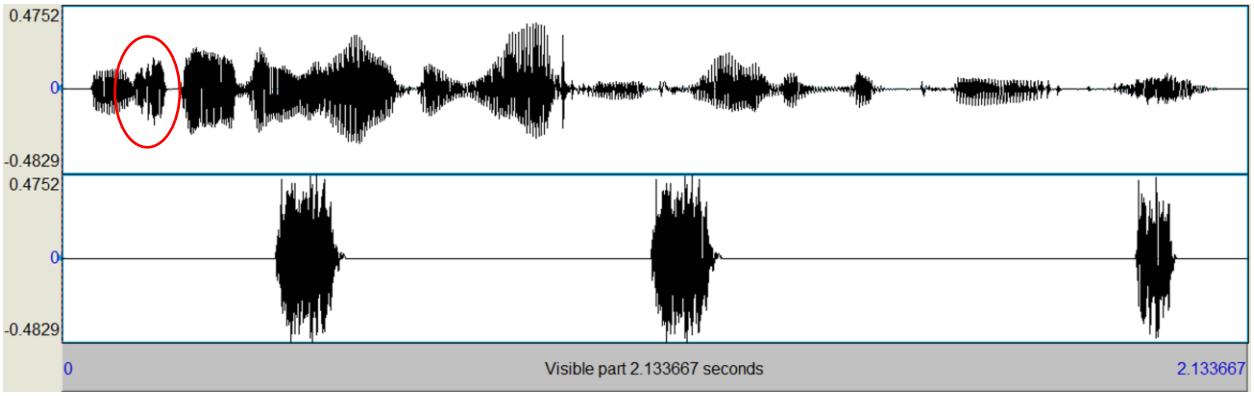
of the /s/ and its corresponding silence in the sentences were time-compressed using Praat to give a more natural sounding sentence. This created a balance where the sentence containing silence did not have an extended pause between the previous word and the base, and the sentence containing the /s/ still had an audible and unambiguous /s/. The compression was applied on a continuum from 50% to 70% in 10% increments, where 50% meant that the /s/ or silence was half the duration of the original. The compression range was the same for early and late position sentences, with comparable means across both (M = 60% early, M = 59% late).

The experimenter made a judgment on how natural the sentences sounded in the timing of the silent gap where the /s/ was placed. If the sentences were judged to be unnatural, some of the additional silence between the /s/ and base onset (or silence and base onset) was cut. The amount of silence cut for early position sentences ranged from 0 to 60 ms (M = 28.75 ms) and the range for late position sentences was 0 to 40 ms (M = 25.94 ms). If the sentences still did not sound natural after that, the original "the" preceding the target word was replaced with a different production of "the" with more clearly articulated phonemes from an unused sentence recording; this occurred for three sentences. The same "the" was used in each sentence that required a replacement. In the late position sentences for the sentences containing *store* and *strip*, the entire target word was replaced with a target word from an unused sentence recording which contained a more clearly articulated stop consonant. A pilot study was conducted which confirmed the naturalness of the stimuli.

In total, there were two sentence frames for each target word, or 32 sentences (16 target words with two frames each). Next, these sentences were converted into stereo, such that there were separate left and right channels for each sentence. This was done twice for each sentence frame, giving 64 sentences. In one sentence, the right channel was completely silenced, leaving

the sentence in the left ear only. This sentence was part of the intact condition because the distance across the head between the /s/ of the target word and the rest of the sentence was separated by 0 degrees. In the other sentence, only the /s/ of the target word was played in the right ear, while the rest of the sentence was in the left ear. This was part of the split condition stimulus because the distance across the head between the /s/ of the target word and the rest of the sentence was separated by 180 degrees. Then an /s/ production from an unused recording was compressed by either 50% or 60% and randomly dispersed within the sentence frames, each containing anywhere from 3 to 5 distractor /s/ sounds in the right ear. Distractor /s/ sounds were included in both early and late sentence positions to mask the presence or absence of the target /s/ in the right channel. An example of a finalized stimulus pair is shown below (Figure 2).

a) He spit away the rotten part of the peach.



b) He bit away the rotten part of the peach.

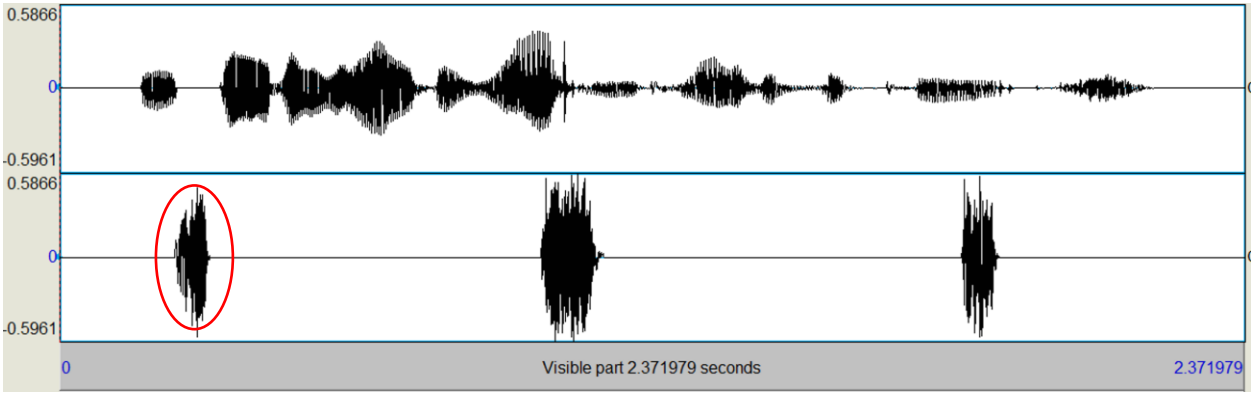


Figure 2: This graph shows acoustic waveforms of the target word spit *sentence pair, which differ only in whether the /s/ of the target word is presented in the left ear (a) or in the right ear (b). In both graphs, the red circle indicates the target word /s/. Time is shown on the x-axis and amplitude of the speech sounds is shown on the y-axis.*

Thirty-two filler sentences were created that did not contain any target word pairs, but simply a sentence in one ear with irrelevant /s/ sounds in the opposite ear. They were constructed to be similar to the target sentences in syllable length, word choice, and distractor /s/ dispersion.

The words in the filler sentences that were in the target word location did not always start with /s/, /b/, /p/, /d/, or /t/. This was to prevent participants from noticing the experimental manipulation. Half of the filler sentences were designed to parallel the early frame stimuli, and half were designed to parallel the late frame stimuli. An example of an early frame parallel filler is "The *pool* is only open in the afternoon," while a late frame parallel filler is "Becca was called in to *work* on the weekend." The early frame fillers started with a one syllable word, as the early frame targets did, and the late frame fillers contained the word "the" or "to" after six to ten syllables in the sentence (M = 7.56 syllables). The filler sentences are listed in Appendix B.

Two stimulus lists were created such that each sentence frame was only present once per list. That is, if the early position sentence for "spike" was present in the split condition in list 1, the early position sentence for "spike" in the intact condition was present in list 2. This prevented participants from hearing the same sentence twice during the experiment. Additionally, if the early position sentence for "spike" was present in the split condition in list 1, the late position sentence for "spike" was present in the intact condition in list 1. Therefore, participants heard each target word twice per list but in different sentence frames: once in the intact condition and once in the split condition.

The stimuli and filler sentences were randomly ordered for each list using Random.org (Haahr, 2019). There were three practice trials and 64 test trials per list, with the test trials broken down further into 32 target trials and 32 filler trials. A sequence of six fillers at the beginning of each test section was added to prevent participants from discovering the experimental manipulation.

Procedure. Participants sat in a sound-attenuated room in front of a computer screen and put on Sony MDR-V900 headphones. They were told that the experiment was a memory test for sentences and that any sounds that were not in the ear of the sentence should be ignored. Participants were assigned to one of the two lists, which alternated between participants. Custom Python code automatically played each audio file in the assigned list with breaks in between each sentence. Participants were instructed to repeat each sentence they heard into a microphone after the whole sentence was played. These responses were recorded using a Sound Blaster Audigy 5/Rx microphone. Participants had the same amount of time it took for the sentence to play to repeat it back, so if the sentence stimulus was 1.5 seconds long, they had 1.5 seconds to speak their response. The experiment lasted for approximately 10 minutes. At the end of the experiment, participants filled out a questionnaire about demographic information. Responses were subsequently transcribed and scored for whether the target word was voiced or voiceless.

Results and Discussion

One participant was excluded because the person did not give any voiced responses and instead responded with only voiceless target words in all trials. The graph of averaged voiced responses for both the intact and split conditions and the early and late positions (N = 27) is shown in Figure 3. I predicted that participants would respond with the voiced target words only in the split condition, and not in the intact condition. Additionally, in the split condition, it was predicted that participants would respond with the voiced target words more in the late condition as opposed to the early condition.

The data were analyzed using a repeated measures ANOVA test in a 2x2 design, comparing voiced responses across the intact and split conditions and the early and late position

sentence frame conditions. In the intact condition, participants rarely responded with the voiced response (M=0.08), compared to the split condition (M=0.61), F (3, 108) = 183.3, $p < 0.001$. This difference indicates that spatial separation was an informative cue to segregation of speech sounds. In other words, people were likely to group sounds together that come from the same location and separate sounds that come from different ones. Performance was about the same for early (M=0.64) and late (M=0.58) positions in the split condition, meaning that participants were reporting the voiced responses at about the same rate regardless of context. Performance was also about the same for early (M=0.07) and late (M=0.08) positions in the intact condition. Taken together, there was no significant main effect of position, F (3, 108) = 0.47, p = 0.50, and no significant interaction between position and spatial separation, F (3, 108) = 0.83, p = 0.37. I predicted that there would be more voiced reports in the split condition for the late positions than the early ones, because selective attention has been shown to increase over time (Best et al., 2008). However, this prediction was not supported by the data; in fact, the opposite is visible in the graph (not significant).

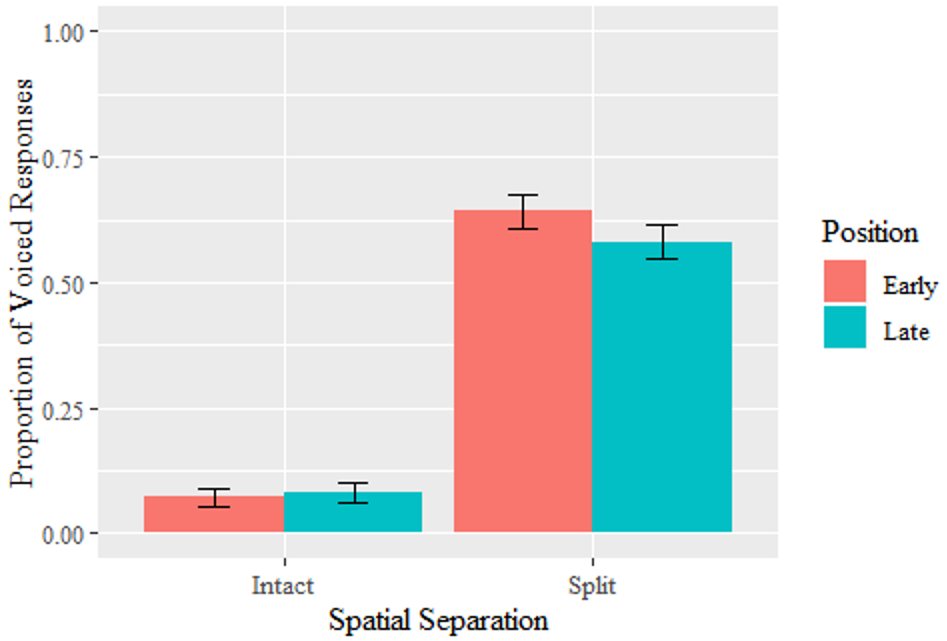


Figure 3: The proportion of voiced ("bike") responses grouped by position and spatial separation are shown here, with spatial separation (split and intact) on the x-axis and the proportion of voiced responses on the y-axis. The standard error is shown in the brackets.

Participants varied drastically in their responses, with the proportion of voiced responses ranging from 0.0 to 0.29 in the intact condition and 0.0 to 1.0 in the split condition. Figure 4 shows performance by participant compared across both positions in each condition. I explored these differences to learn how the pattern of results across conditions differed among individuals. Most participants increased in their proportion of voiced responses from intact to split stimuli (N=27), though a few remained the same despite spatial separation, like participant 3 in the late condition, or decreased, like participant 8 in the early condition. This individual variation suggests that the integration of the /s/ with the base word is not the same for every person. This could be attributed to differences in selective attention, as a result of experience with sustained attention to sounds or language. This listening experience might affect whether they are likely to

integrate sounds or segregate them. These individual differences can be characterized as stable because the participants behaved consistently across the experiment, as shown in Figure 5 by the positive correlation between how participants performed in the early and late position, $r(25) = 0.52$, $p = 0.006$.

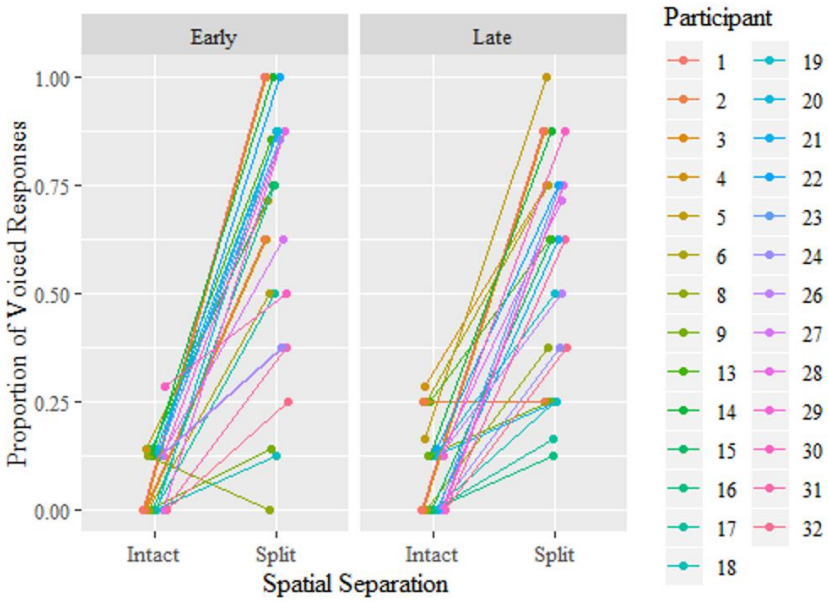


Figure 4: The proportion of voiced (bike) responses for each position and spatial separation are shown here for each individual, with spatial separation (split and intact) on the x-axis and the proportion of voiced (bike) responses on the y-axis. There are separate graphs for the early and late positions. The lines show the relationship of the proportion of voiced (bike) responses per participant for the intact and split conditions in both the early and late positions.

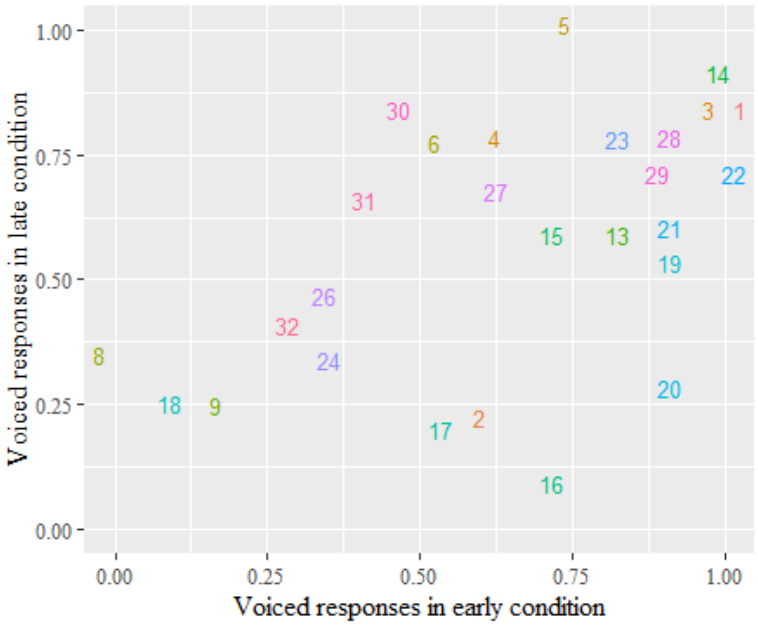


Figure 5: Individual participants' voiced responses in the early position (x-axis) versus the late position (y-axis) are plotted above. Each number represents one participant.

Two participants (participants 3 and 8) were selected that did not show the expected pattern between the intact and split conditions. Although both of these participants behaved consistently across the experiment, as shown in Figure 5, they did not show an increase in voiced responses from the intact to the split condition of the late position sentences. Therefore, their questionnaires were examined to see whether they reported any sustained language or close listening experience. Participant 3 self-reported having college experience with Spanish and speaking it at a fluency level of 3 out of 5, with 1 being the least fluent and 5 being the most fluent. Additionally, participant 8 self-reported having no college experience with a language and stated they could speak Spanish at a level of 2 out of 5. In contrast to participants 3 and 8, participant 31 showed the expected pattern, but had no experience with foreign language or with speech sound discrimination. Therefore, performance on this task might be due to individual

differences that are not related to prior experience with learning a new language or selectively listening to sounds.

Average performance was also examined from the first half to the second half of the trials in order to investigate whether the overall proportions of voiced responses in each sentence condition was consistent over time. Ideally, participants would show the same rate of voiced responses over the entire experiment because of the balanced lists and repetitive task. If participants' voiced reports were inconsistent across the experiment, it could indicate that they had become familiarized with the experimental manipulation. It could also indicate a list imbalance. However, if their voiced reports decreased over the blocks, it could either be a sign of fatigue in addition to (or instead of) a list imbalance. Split condition voiced responses decreased from the first to the second half in both the early and late conditions (early condition: 0.78 first half, 0.53 second half; late condition: 0.69 first half, 0.49 second half).

This difference from performance in the first half to the second half led to a closer examination of the lists for the underlying cause. The initial randomization process created lists where several target trials were presented in succession, followed by several filler trials in succession. This also affected the number of each type of stimulus in each half; for example, in the second list there were 19 filler trials and 13 target trials in the first half, resulting in 13 filler trials and 19 target trials being present in the second half. The way the lists were randomized (i.e. the sequential runs of target trials) might have led to habituation to the experimental manipulation and an increased tendency to integrate every sound over time, leading to a decrease in voiced responses. In fact, the results of an independent samples t-test measuring the differences between means of voiced responses to each of the lists indicate that the orders of the two lists were not evenly randomized, $t(1676) = -4.37$, $p < 0.001$. A solution for this problem

would be to reorganize the lists into a pseudo-random format to counterbalance all the types of stimuli. This would prevent participants from noticing patterns in the target stimuli and instead allow them to experience a balanced experiment, where each stimulus cannot be directly compared with the stimuli before and after it.

Overall, there was no main effect of temporal position on participants' voiced responses, suggesting that within a sentence context, participants are no more likely to integrate spatially separated extraneous sounds at the beginning than at the end. This contrasts with the results of Best et al. (2008), who found that selective attention strengthens over time, causing increased segregation of extraneous sounds from the attended speech stream. Instead, the results of the current experiment indicate that speech sounds are equally integrated at all points within a speech stream, suggesting that when the streams consist of speech sounds with context, they do not behave like other auditory streams (i.e. those of environmental sounds, containing no linguistic context). While Best et al.'s experiment did have context in terms of the phonemes that came together to form each number word, it did not have the semantic context that is provided by a sentence, where it is possible to predict upcoming words or phrases. Generalizable interpretations of daily speech processing thus could not be drawn from that experiment as readily, because semantic context may contribute to the interpretation of a complex auditory scene. It would seem that the introduction of semantic context would increase cohesion of the words within the stream over time and show a stronger effect than the one seen in Best et al.'s experiment. However, this was not the case in the current experiment, indicating that semantic context may help the listener maintain a constant level of selective attention while listening to a speech stream.

Additionally, there was a main effect of spatial separation, indicating that it is an informative cue for speech segregation. This use of an acoustic cue for language perception could indicate that the auditory environment surrounding language is important in deciphering which speech sounds are related. Bregman's (1990) theory of Auditory Scene Analysis states that sounds that are grouped together primitively are organized regularly in the environment with aligned acoustic properties. The results of this experiment lead to the conclusion that spatial location is a cue for continuity of speech sounds, supporting Bregman's theory. Because the sounds that were presented in the opposite ear from the rest of the sentence were more often perceptually separated, it is possible to extrapolate that speech sounds are more likely to be grouped together when they come from the same location. Finally, I found persistent individual variation, which indicates that the task of sound integration and segregation may be facilitated by experience or some other factor (individual differences in selective attention). However, as described above, these results could be due to confounding variables such as experimental habituation and list imbalances.

Experiment 2

Experiment 2 was a replication of Experiment 1 with refinements of the methodology in order to rule out patterns in the results that could have been caused by factors other than the manipulated variables. This would strengthen the claim of no effect of position that was proposed in Experiment 1. Sentence frames that participants had difficulty with were replaced, potential order effects within the stimulus lists were removed, and foil trials were added. All of these changes were made in order to improve the experiment quality.

Method

Many of the details of the experiment were the same as Experiment 1. Only differences are described. Firstly, two sentences were replaced due to participants responding with the voiceless target word in the early position of the intact condition over half the time (0.54 for “spill” and 0.6 for “spare”). Additionally, breaks were added every 19 trials, giving a total of three breaks, in order to mitigate concerns about fatigue. The length of each break was determined by each individual participant.

Participants. Thirty-one Ohio State University undergraduate students (12 females, M = 20.65 years, range = 18–38 years) participated in the experiment and received course credit. All of the participants were self-reported native English speakers with normal hearing.

Stimulus creation. In order to determine which sentence frames needed to be replaced, participant responses in experiment 1 to the target words in the intact condition were compiled. These responses were then counted, and a proportion for each sentence was calculated that

showed the number of devoiced responses out of the total responses. If the proportion of devoiced responses was lower than 0.5, the sentence was replaced, because this meant that participants were biased to report the voiceless word within the sentence by its context more than half of the time. In addition to the replacement of some sentences, changes were made to the list order to better balance the sentences between fillers and targets, and breaks were implemented throughout the experiment.

Foil trials were created to detect participants who failed to follow instructions. Foil trials were designed to be similar to test trials in terms of sentence structure and location of the target word. For example, if participants heard the sentence, “The bat flew into the barn after sunset,” in their left ear and a corresponding /s/ in the right ear immediately before the word “bat,” they might have a strategy to report the sentence, “The spat flew into the barn after sunset.” If they did so, they would be excluded from the results. Therefore, if participants were strategically grouping the /s/ with the base during the foil trials, it would result in a sentence that does not make sense.

Foil sentences were created in a similar manner to the target sentences, such that they contained a target word which could have an /s/ added onto it to make another real word (for example, “pin” can have an /s/ added to make “spin”). However, only the target word without the /s/ made sense within the sentence context. For example, in the sentence “You can tell her name by the pin on her shirt,” adding an /s/ onto “pin” to make “spin” would result in an illogical sentence. The same female native-English speaker recorded the sentences with the /s/ included on the target word, and this /s/ was moved using Praat (Boersma & Weenink, 2019) to the opposite ear from the rest of the sentence, keeping its natural position in the word (before the base in “spin”). Additional distractor /s/ sounds were included in each foil to match the other

target and filler sentences. Foil sentences were marked with either the early or late position because they were designed to parallel the target sentences in structure. In other words, the early position fillers started with “the” or “a” followed by a one syllable word, and the late position fillers had “to” or “the” several syllables into the sentence followed by a one syllable word. Twelve foil sentences were created in total, which came from six key words. For example, the word “spat” produced one foil sentence where the target was “bat” and one foil sentence where the target was “pat.” Three foil sentences were evenly dispersed into each of the four blocks in the experiment such that all twelve foil sentences were included in each list. Foils are listed in Appendix C.

The sentence frames that replaced those in Experiment 1, including their silence compression and removal information, are listed in Table 1. These are new frames that were recorded for stimuli with which participants consistently made errors. The same compression level as the silence compression was applied to the corresponding /s/ in the opposite channel. Additionally, the same amount of silence that was removed in the sentence channel was removed from the opposite channel directly before the target /s/ production. All of the new frames contain the same number of syllables as the original sentences from Experiment 1.

word	position	sentence	silence compression	silence removal (ms)
stew	late	She got me to lick the X off the grass.	50%	0
spill	early	The X would be expensive to take care of.	50%	30
strip	late	My new barber tells me not to X my hair dry.	50%	0
spare	late	My closet has room for a X costume.	50%	30

Table 1: The replacement sentences for the original stimulus items from Experiment 1 are listed here. Silence compression denotes the amount that the silence left behind when the /s/ of the target word was lateralized to the opposite channel was compressed. Silence removal refers to the amount of silence that was removed between the "the" and the base word in the sentence channel.

In Experiment 1, all of the sentences were played in the left ear with the distractor /s/ sounds in the right. In Experiment 2, each list had a version where the sentence was played in the left ear and a variant where it was played in the right ear to counterbalance the ear of presentation. That is, from the original Lists 1 and 2 in Experiment 1, two more lists were created for Experiment 2 such that there were List 1-left, List 1-right, List 2-left, and List 2-right. List 1-left and List 1-right contained the same order of items; the only difference was the ear of sentence presentation. The same was true for List 2-left and List 2-right. Each participant was assigned to one of those four lists. There were five practice trials and 81 test trials per list, broken down further into 32 target trials, 32 filler trials, and 12 foil trials. The filler trials were the same as those in Experiment 1. The lists were pseudo-randomized so that there were no runs of multiple trials of the same type, except at the beginning where a long run of fillers was included to make the experimental manipulation less salient at experiment onset (like in Experiment 1).

Procedure. The procedure for this experiment was identical to that of Experiment 1, except participants were randomly assigned to one of the four lists mentioned above. An additional difference was that there were three breaks in the experiment. The length of those breaks was determined by each participant.

Results and Discussion

One participant (participant 7) was excluded for not following instructions. No participants were excluded based on incorrect foil answers, as they were very rare overall.

The graph of the averaged voiced responses for both the intact and split conditions and the early and late positions of the replication experiment (N = 30). is shown in Figure 6. The data were analyzed using a repeated measures ANOVA test in a 2x2 design, comparing voiced responses across the intact and split conditions and the early and late position sentence frame conditions. Similar to Experiment 1, participants rarely responded with the voiced response in the intact conditions (M=0.04), compared to the split conditions (M=0.71), F (3, 120) = 282.65, p < 0.001. This result was a direct replication from the first experiment, reinforcing that spatial separation is an informative cue for the segregation of speech sounds. Also like Experiment 1, performance was similar for early (M=0.75) and late (M=0.67) positions in the split condition, indicating that context did not matter when it came to the rate of voiced responses. Performance was also similar for early (M=0.04) and late (M=0.05) positions in the intact condition. Taken together, there was no significant main effect of position, F (3, 120) = 0.47, p = 0.49, and no significant interaction between position and spatial separation, F (3, 120) = 0.63, p = 0.43.

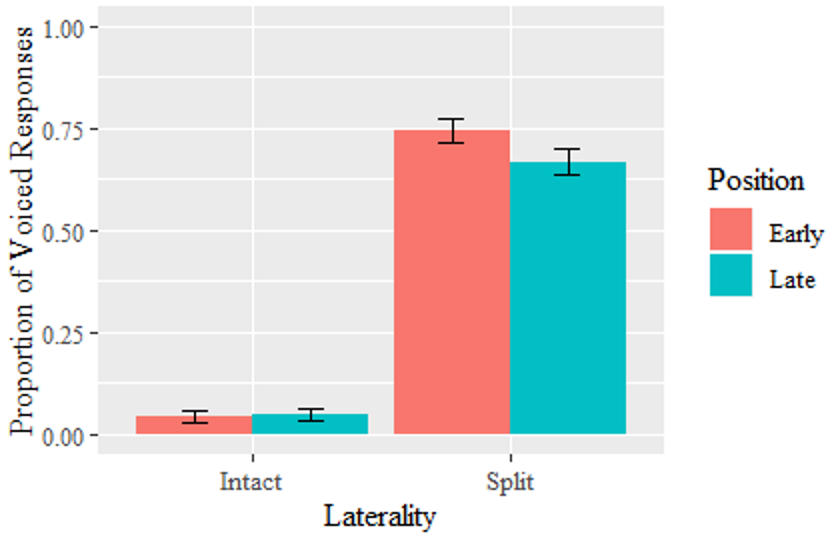


Figure 6: The proportion of voiced responses for each position and spatial separation are shown here, with spatial separation on the x-axis, the proportion of voiced responses on the y-axis, and standard error brackets.

Participant variability in this experiment was also high, with the proportion of voiced responses ranging from 0.0 to 0.29 in the intact condition and 0.0 to 1.0 in the split condition. Figure 7 shows performance by participant compared across both positions in each condition. However, there were not the same types of outliers as in Experiment 1. For example, no participants reported fewer voiced responses in the split condition than in the intact condition in this experiment. One participant showed no voiced responses at all in the late condition (participant 3), but overall, participants showed the expected trend of fewer voiced responses in the intact condition as compared to the split condition. The diminishing of outliers could be due to the balancing of the lists, with fewer participants figuring out the experimental manipulation.

These were the same results that were found in Experiment 1, indicative of similar conclusions. Because there had been several adjustments to the experimental method and the

results remained the same, the conclusions from the original experiment were strengthened. The most notable of these conclusions were that spatial separation is a reliable cue to indicate which speech sounds should be integrated together, and that variation in level of integration exists among participants.

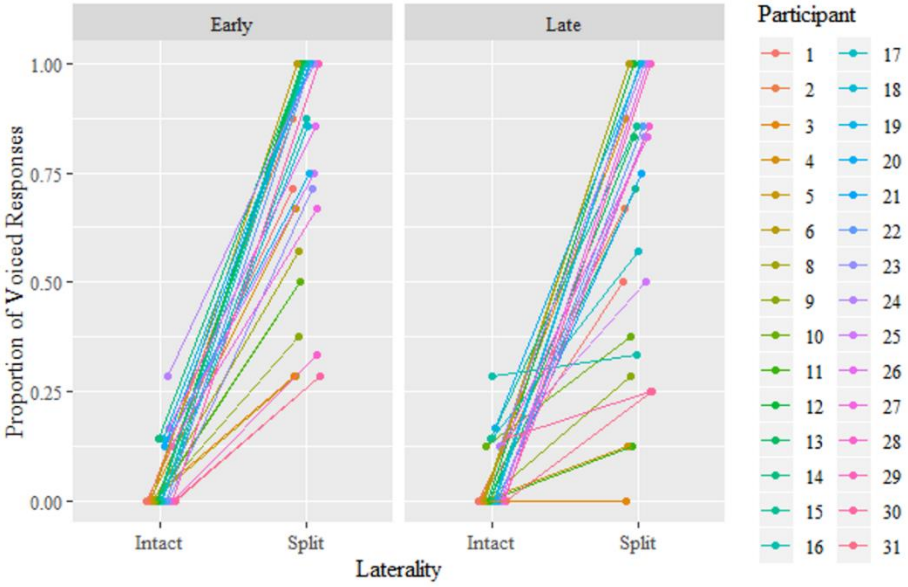


Figure 7: The lines show the relationship of the proportion of voiced (bike) responses per participant for the intact and split conditions in both the early and late positions. There are separate graphs for the early and late positions.

There was not a significant difference in comparison of lists 1-left and 2-left, $t(1703) = 1.61$, $p = 0.11$, or in comparison of lists 1-right and 2-right, $t(1702) = 1.63$, $p = 0.1$. Only lists with the same ear of presentation were directly compared. There was no significance when comparing lists that only differed in presentation order.

Analysis of this experiment over blocks still showed a decrease in voiced responses in the split condition over time, despite the addition of breaks and pseudo-randomization of the lists.

Over each of the four blocks (each with 19 trials), the proportion of voiced responses decreased steadily in the split condition, as shown by the graph of the intact and split conditions and the early and late positions split by block in Figure 8. This could mean that selective attention becomes less refined over time and thus allows for increased integration over the course of the experiment.

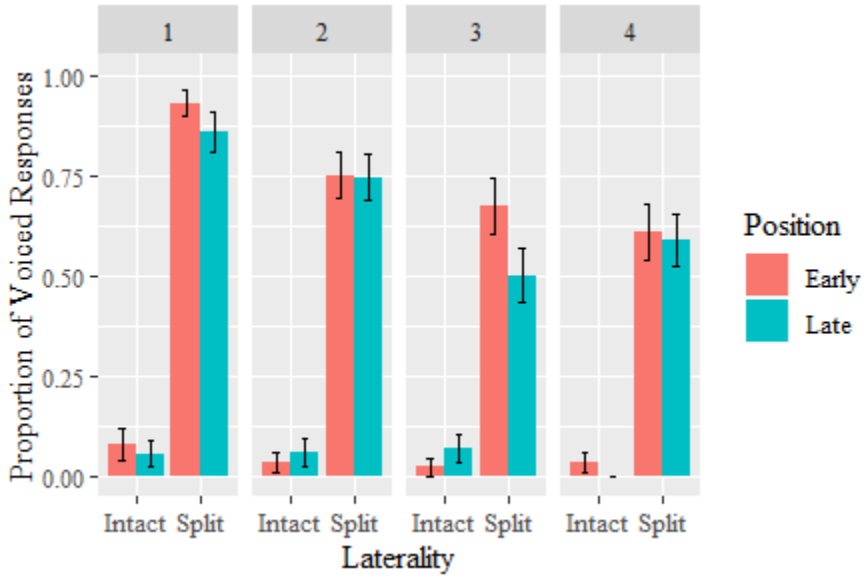


Figure 8: The proportion of voiced responses for each laterality and position are shown here, separated into four blocks where each block is represented by a different graph (numbered 1-4). Laterality is shown on the x-axis, while the proportion of the voiced responses averaged over all of the participants is shown on the y-axis. The standard error bars are also shown here.

It was predicted that speech sounds would become more cohesive as a sentence goes on, based on the results of Best et al.'s (2008) study which found that selective attention is refined over time. The results instead show no effect of position (where the target word occurs in the sentence) on amount of voiced responses. This implies that speech sounds are equally integrated

at all points within a speech stream. An increase in selective attention might have been achieved in these experiments if there was a more direct replication of the cocktail party effect, as was created by Best et al. (2008). The multiple simultaneous talkers used in Best et al.'s experiment created a more complex auditory scene that could have led to participants' increase in selective attention over time. The present experiments only presented one sentence at a time, potentially eliminating the need for a refinement of selective attention.

General Discussion

The current study tested the proposal made by Best et al. (2008) that continuity from a single location in space causes speech to increase in coherence over time. The key experimental finding was that there was no difference between the responses to the early and late positions in the split condition. This contrasted with the prediction that there would be more voiced responses in the late condition than the early condition, which originated from analysis of the results of Best et al. Despite the results between the early and late positions being unexpected, the expected results between the intact and split conditions were produced via the spatial separation manipulation. Surprisingly, there was a wide variety of individual differences, which were stable and persistent across conditions.

The difference that was expected between the early and the late position of the split stimuli was not found. This indicates that the amount of context provided before the key word in the sentence (one word or many words) was irrelevant in terms of whether the spatially separated sound was perceptually attached to the base or not. All split sentences were simply perceived by the participants as separate from the base about 75% of the time, regardless of whether they were in the early or the late position. Best et al. (2008) found that selective attention increases over time the longer a stream of words progresses. What can be gleaned from the results of these studies is that selective attention might not be strengthened over time when it comes to a linguistic context. That is, people might just have a baseline level of selective attention that does not get stronger or weaker as sentences progress. This drastic difference from Best et al. could be due to the minimally complex auditory scene provided in this experiment.

The spatial separation manipulation produced the expected result, meaning that sounds that were spatially separated from the main sentence were kept perceptually apart from the

sentence most of the time. In this way, spatial cues were used to influence which sounds were judged as relevant speech sounds and which were considered irrelevant. This expected result was not produced in every case, however, because about 25% of responses remained voiceless when the /s/ and base were split. This result could be due to individual differences; not all participants responded in a similar way to each stimulus.

Participants reported varying levels of experience with learning foreign languages and carefully discriminating speech sounds, potentially causing them to perform in different ways in this experiment. However, people who did not have a lot of experience with carefully discriminating sounds or learning new languages still showed the pattern that was expected. In Experiment 1, 33% of participants behaved as predicted, and of those, 44% reported having foreign language skills. Similarly, in Experiment 2, 27% of participants showed the expected pattern of responses. Of them, the percentage of participants reporting foreign language skill was 38%. Therefore, it does not seem predictable from the prior experiences measured in this study how participants are going to behave in selective attention-based tasks. Bregman (1990) does not discuss individual differences as a factor in primitive segregation because he states that the features by which speech sounds are categorized are objective acoustic categories. These can be perceived by any human with normal hearing.

The conclusions reached by Best et al. (2008), that selective attention is strengthened over time as a stream progresses, were not supported by the results of both experiments. More voiced reports were expected in the late position of the split condition sentences compared to the early position in order to support this claim. However, the data do not show this difference. This could be due to the differences in experimental manipulation between the studies outlined in this paper and the study completed by Best et al. The stimuli used in the present experiments were

spoken sentences with linguistic context, while the stimuli in Best et al.'s experiments were spoken strings of numbers. Participants in Best et al.'s experiment repeated strings of numbers that they heard from loudspeakers at different spatial locations, while participants in these experiments repeated sentences that consistently came from one ear with distracting sounds in the opposite ear. Out of all of these, it is most likely the different results in the current experiments came from the lack of presentation of multiple simultaneous auditory streams. Because the current experiments only involved one sentence being presented at a time, and Best et al. had multiple strings of numbers presented at once, the latter is a complex auditory scene that more closely resembles the cocktail party problem. It is possible that the refinement of selective attention over time only occurs in a sufficiently complex environment, which explains why this refinement was not attested in the current experiments. Additionally, while Bregman (1990) did address the role of conscious attention in schema-based segregation, he did not mention the enhancement of attention (whether conscious or not) over time. Therefore, the results of the experiments performed here seem to support his theory of Auditory Scene Analysis because there was no apparent increase in segregation, and by extension no increase in selective attention.

Bregman's (1990) concept of primitive segregation states that acoustic cues such as spatial separation are enough to indicate that two sounds should not be grouped together. The results therefore somewhat support his theory of Auditory Scene Analysis because the sounds presented in the opposite ear from the rest of the sentence were segregated at an average rate of 66% in the split condition across both positions and experiments. While some participants performed close to the average, others integrated the /s/ with the base every time, regardless of whether the word was in an early or late position. If Bregman's (1990) claim that spatial

separation is an informative cue for primitive segregation was fully supported by the data, I would have expected 100% segregation across the experiment. However, early and late positions had different rates of segregation in the split condition (*M*=0.7 early, *M*=0.63 late), and people differed in their amount of segregation overall. Therefore, there was a wide range of individual differences across the experiment.

In the context of speech, some people in the experiment interpreted ambiguous sounds as speech. Even though participants were told that the sounds in the opposite ear were irrelevant, several of them did mention in the questionnaire that the sounds presented in the opposite ear from the sentence sounded like speech sounds. For example, in Experiment 1, 26% of the participants noted that the sounds in the opposite ear from the sentence sounded like /s/ sounds. Those who noticed this mentioned that it affected their ability to repeat back words when it changed their meaning (for example, when deciding whether they heard the word “bike” versus the word “spike”). 33% of the Experiment 1 participants stated that they believed the experimental goal was to investigate how extraneous noise affected sentence perception, and 27% of participants stated the same thing in Experiment 2. In Experiment 1, the participants who characterized the experimental goal in this way reported more voiced responses to the target words in the split condition/early position sentences than the other 67% of participants. A difference between those two groups of participants was also seen in Experiment 2, with more voiced responses to the target words in the split condition/late position sentences reported by the 27% of participants who accurately described the experimental goal.

Future experiments would involve separating the sentences and /s/ sounds by 150 degrees instead of 180 (making the perceptual segregation harder) to see if that draws out the slight difference seen in the early and late position sentences in the split condition. If it were possible to

find significantly more voiced responses in the late position sentences rather than the early position ones, this would provide support for the conclusions drawn by Best et al. (2008). There was no difference in either of the experiments when examining that comparison, although a difference would have been predicted by Best et al. By separating the sentences from the /s/ sounds by 150 degrees, it would be slightly harder for participants to distinguish which sounds are coming from which ear. If there has been a ceiling effect for voiced responses in the split condition when separating the /s/ sounds by 180 degrees, bringing the /s/ and sentence closer together would hopefully prevent participants from hitting that ceiling and instead showing a more distinct difference between positions.

Speech sounds that are spatially separated from the rest of a speech stream are resistant to being grouped in with the rest of the speech stream, as was theorized by Bregman (1990). Additionally, selective attention to a particular speech stream does not seem to increase as a sentence progresses, contrary to what would have been predicted by Best et al. (2008). A complex auditory scene might lead to refinements in selective attention that are not seen in an auditory environment involving only one talker with minimal extraneous noise. These results provide an insight into speech perception because they address how speech sound processing operates differently in different environments.

References

Best, V., Ozmeral, E. J., Kopčo, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences*, *105*(35), 13174-13178. https://doi.org/10.1073/pnas.0803718105

Boersma, P., & Weenink, D. (2019). *Praat: doing phonetics by computer [Computer program]*. Retrieved from https://www.fon.hum.uva.nl/praat/

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. MIT press. https://doi.org/10.7551/mitpress/1486.001.0001

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, *25*(5), 975-979. https://doi.org/10.1121/1.1907229

Haahr, M. (2019). *RANDOM.ORG: True Random Number Service.* Retrieved from https://www.random.org/lists

Lany. (2015). *Back of head* [PNG file]. Retrieved from http://dynamicpickaxe.com/back-of-head-clipart.html#

Appendix A: Experiment 1 Target Sentences

The target sentences used in Experiment 1 are listed in the table. The X indicates the location at which the target word (with or without the /s/) could be inserted.

Target Word	Base	Early Position	Late Position
space	base	The X was big enough for the both of us.	We tried to take over the X yesterday.
spare	bare/bear	The X shelf was in the red cabinet.	For Halloween I wore the X costume.
spark	bark	The X flew toward us when the fireworks hit the trees.	Maria was so startled by the X that she jumped.
spear	beer	The X was made before our son Evan was born.	It was hard to keep the X away from the kids.
spike	bike	The X hurt his foot when Joe stepped in the yard.	Sam had to look out for the X in the road.
spill	bill	The X was larger than we had expected.	I tried to avoid the X on the table.
spit	bit	He X away the rotten part of the peach.	The fight was so fierce that he X back his words.
stab	dab	They X the turkey meat and put it in the hot pan.	It felt like someone was trying to X at the wound.
start	dart	My X did not earn me any points in the first round.	The announcer told them to X around the corner.
steel/steal	deal	The X was key for the construction of the building.	She quietly told Michael to X the blackjack cards.
stare	dare	The X was intimidating enough to scare Jill.	I could not believe the X that Cam gave me last night.
steer	deer	The X bolted away from the farmer.	They slowly tried to feed the X a pear.
stew	dew	The X got on my expensive new shoes.	She dared me to lick the X off the grass.
store	door	The X is closing after the last person leaves.	Lou will finally open the X tomorrow.
strain	drain	We X the pasta after it has finished cooking.	Dave knew the cause of the X on the economy.
strip	drip	The X of wax was sticking to my hairy leg.	My stylist did not want me to X my hair dry.

Appendix B: Filler Sentences

The filler sentences used in both Experiments 1 and 2 are listed here. The italicized word in each sentence shows the word in a parallel location to the target words in those sentences.

Early Position	Late Position
The *chef* had been making his special dish for years.	Ryan called his dad to *ask* for his advice.
A *piece* of cherry pie was nice after the meal.	Becca was called in to *work* on the weekend.
The *pool* is only open in the afternoon.	After I napped I started to *feel* better.
A *work* trip took Dan to Tokyo last summer.	Nick went to the pet store to *buy* a goldfish.
The *clock* in the living room has been wrong for years.	I took a vacation to *get* a break from school.
The *park* had just gotten a new swing set put in.	She is training to *run* a marathon next spring.
The *dog* was barking loudly enough to wake us up.	The annoying cat caused me to *trip* on the stairs.
The *book* was popular thanks to its famous author.	Charlie does not like the *songs* on the radio.
The *drawer* was full of mechanical pencils and pens.	Tim asked the waiter to *box* up the leftovers.
The *jazz* music was relaxing after a long day.	I wanted my best friend to *stay* another week.
The *girl* showed off her singing in the school talent show.	Ashley was not brave enough to *try* a new food.
A *breeze* blew in through the open window.	Some people think they need to *have* coffee daily.
The *woods* were full of wild animals.	I love to play the *flute* in my free time.
The *fish* were jumping out of the river.	A knock at the door caused my dog to *jump* off the couch.
The *dark* clouds signaled a storm heading our way.	My mom promised to *take* us to the zoo.
The *man* helped his girlfriend move to a new house.	George never learned how to *swim* the backstroke.

Appendix C: Foil Sentences

The foil words with their sentence frames are listed below.

Correct	Incorrect	Early Position	Correct	Incorrect	Late Position
punk	spunk	A X vandalized the store last weekend.	bunk	spunk	She was sick of climbing up the X bed each day.
toe	stow	My X was poking out of my old sock.	dough	stow	We needed to wait for the X to double in size.
tile	style	The X in the kitchen was refinished.	dial	style	He was feeling too nervous to X the number.
bat	spat	The X flew into the barn after sunset.	bat	spat	The owner said we could try to X the small dog.
beak	speak	The X of his bird poked around the corner.	peek	speak	My sister told me not to X out the window.
bin	spin	The X was just big enough to fit all my clothes.	pin	spin	You can tell her name by the X on her shirt.